IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

Li *et al.*

Appl. No. 09/720,086

102(e): July 23, 2001

For: ***De Novo* DNA Cytosine Methyltransferase Genes, Polypeptides and Uses Thereof**

Confirmation No.: 6968

Art Unit: 1642

Examiner: Harris, A. M.

Atty. Docket: 0609.4560002/KRM/DJN

Declaration Under 37 C.F.R. § 1.132 of En Li, Ph.D.

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

I, the undersigned, En Li, Ph.D., residing at 45 Hinckley Road, Newton, Massachusetts, 02168, declare and state as follows:

1. I am a co-inventor of the above-captioned patent application.

2. I am currently employed by Novartis Institutes for Biomedical Research as Vice President & Global Head, Animal Models of Disease and Epigenetics Program. Prior to my current employment, I was an Associate Professor of Medicine at Harvard Medical School and directed a laboratory in the Cardiovascular Research Center at the Massachusetts General Hospital from January 1993 to April 2003, where I conducted and supervised research in the field of mouse genetics and developmental biology.

3. A current *curriculum vitae* is appended hereto as EXHIBIT A.

4. I have reviewed the above-captioned patent application and the Office Action dated June 6, 2005. I have also reviewed the sequence listing as filed and the sequence listing as amended on July 23, 2001. I have also reviewed the claims of the captioned patent application.

5. I have been informed that the Examiner has not granted priority to the earlier filed patent applications because there is insufficient proof that the coding regions of currently amended **SEQ ID NOS:1** and **2** are the same as those listed in the priority documents, viz., the mouse Dnmt3a and Dnmt3b cDNA clones encoding the coding regions of **SEQ ID NOS:1** and **2**, respectively, that were deposited with the American Type Culture Collection (ATCC), 10801 University Boulevard, Manassas, Virginia 20110-2209, USA. Sequences harboring the coding regions of **SEQ ID NOS: 1** and **2**, respectively, were

deposited with the ATCC on June 16, 1998, and assigned ATCC Deposit Nos. 209933 and 209934, respectively. The deposit date of June 16, 1998 was prior to the filing date of the first provisional application, App. No. 60/090,906, filed June 25, 1998, the benefit of which is claimed. The '906 application includes the sequence information and references the deposits of the sequenced material on page 15, lines 26, through page 16, line 2, of the specification.

6. In November 2004, Applicants had samples withdrawn of the mouse Dnmt3a and Dnmt3b cDNA clones contained within ATCC Deposit Nos. 209933 and 209934, respectively. At the Applicants request, Kenneth D. Bloch, M.D., a faculty member in the Cardiovascular Research Center at the Massachusetts General Hospital and an experienced DNA sequencer, sequenced nucleotides that spanned the coding regions of mouse Dnmt3a and Dnmt3b in the deposited cDNA. A nucleotide alignment that spans the coding regions of sequenced mouse Dnmt3a cDNA clone contained in ATCC Deposit No. 209933 and currently amended **SEQ ID NO:1** is shown in EXHIBIT B. A nucleotide alignment that spans the coding regions of sequenced mouse Dnmt3b cDNA clone contained in ATCC Deposit No. 209934 and currently amended **SEQ ID NO:2** is shown in EXHIBIT C.

7. The amendment to the sequence listing, which was filed on July 23, 2001, corrected six nucleotides in the coding sequence of **SEQ ID NO:1** (see the bolded nucleotides at positions 516, 843, 1036, 1110, 1116 and 1726 in EXHIBIT B) and two nucleotides in the coding sequence of **SEQ ID NO:2** (see the bolded nucleotides at positions 918 and 920 in EXHIBIT C).

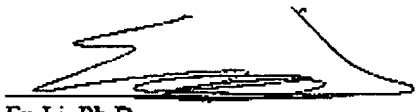
8. The deposited clones recited in ¶¶5 and 6, above (*i.e.*, ATCC Deposit Nos. 209933 and 209934) are the same as the deposited clones recited in the above-captioned application. The coding sequence of ATCC Deposit No. 209933 is currently believed to be the same as the coding sequence of currently amended **SEQ ID NO:1**. The coding sequence of ATCC Deposit No. 209934 is currently believed to be the same as the coding sequence of currently amended **SEQ ID NO:2**.

9. It is well known that sequencing errors are a common problem in Molecular Biology. *See, e.g.*, Peter Richterich, Estimation of Errors in 'Raw' DNA Sequences: A Validation Study, 8 *Genome Research* 251-59 (1998)(EXHIBIT D). I believe that one skilled in the art would have sequenced the deposited material and recognized the sequencing errors in the coding region. I believe that the correct mouse Dnmt3a and Dnmt3b coding sequences are inherent to the ATCC deposited clones, ATCC Deposit Nos. 209933 and 209934, respectively, which were deposited prior to the filing of App. No. 60/090,906, filed June 25, 1998, the benefit of which is claimed.

10. Accordingly, based on the above, I believe that Applicants are entitled to the June 25, 1998 filing date for the coding sequences of mouse Dnmt3a and Dnmt3b contained within ATCC Deposit Nos. 209933 and 209934, respectively.

11. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the present patent application or any patent issued thereon.

Respectfully submitted,



En Li, Ph.D.

Date: 11/7/05

304890v1

CURRICULUM VITAE

En Li, Ph.D.

Vice President & Global Head
Animal Models of Disease
Novartis Institute for Biomedical Research, Inc.
250 Mass Ave. Cambridge, MA 02139
Tel: 617-871-7072
Fax. 617-871-7263
Email. en.li@novartis.com

Education:

1984	B.Sc. in Biochemistry. Peking University, Beijing, China
1992	Ph.D. Massachusetts Institute of Technology (Biology) (Advisor: Prof. Rudolf Jaenisch)

Professional Experience:

1993-1996	Principal Investigator, Cardiovascular Research Center, Massachusetts General Hospital Instructor, Department of Medicine, Harvard Medical School
1996-2000	Assistant Professor, Department of Medicine, Harvard Medical School
2000-2003	Associate Professor, Department of Medicine, Harvard Medical School
2001-2003	Guest Professor, Beijing University, Health Science Center
2003-	VP & Global Head, Models of Disease Center, Epigenetics Program Novartis Institute for Biomedical Research.

Review and editorial board

1999-	External Grant Reviewer, Human Frontier Science Program
1999-	External Grant Reviewer, NIH, NICHD
1999-	Member of the Advisory Board. Journal of Biochemistry
2000-	External Grant Reviewer, NIH, NIA
2000-	External Grant Reviewer, NSF
2000-	Mail Reviewer, Wellcome Trust
2003-	Member of the Advisory Board. China Science Reports
2004	External Grant Reviewer, Chinese Natural Science Foundation.

1993-	Ad Hoc Reviewer for the following journals Nature, Science, Cell, Nat Genet, Genes Dev, Trends Genet, Development, Mol Cell Biol, PNAS, Human Mol. Genet., Dev. Biol., Mech. Dev., J. Cell Biol., Dev. Dyn., Gene, Genomics, Nucl Acid Res, Mammalian Genome, etc.

Professional Societies

1990-	American Association for the Advancement of Science, Member

1994- DNA Methylation Society, Member
1999- Ray Wu Society, Member

Invited Presentations (Since 2003-)

2003 Invited speaker, Gordon Research Conference - Cancer Genetics and Epigenetics
2003 Session chair, Keystone Symposium – Chromatin (Big Sky, Montana)
2003 Invited speaker, Annual HUGO meeting at Cancun, Mexico
2003 Invited speaker, Gordon Research Conference – Epigenetics
2004 Invited speaker, the 2nd Annual CDB symposium, Kobe, Japan
2004 Invited speaker, 2nd Weiseenburg Symposium on DNA methylation – an important genetics signals. Weissenburg, Germany
2004 Session Chair, on genomic imprinting. 10th SCBA International Symposiums, Beijing, China
2004 Invited speaker, Genomic imprinting workshop in Montpellier, France
2005 Vice Chair, Gordon Research Conference 'Cancer Genetics and Epigenetics'

Publication:

1. Zijlstra M, Li E, Sajjadi F, Subramani S, Jaenisch R. Germ-line transmission of a disrupted b2-microglobulin gene produced by homologous recombination in embryonic stem cells. Nature 1989; 342: 435-8.

2. Lee K, Li E, Huber J, Landis S, Sharpe A, Chao MV, Jaenisch R. Targeted mutation of the p75 low affinity NGF receptor gene leads to deficits in the peripheral sensory nervous system. Cell 1992; 69: 737-49.

3. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 1992; 69: 915-26.

4. Li E, Sucov HM, Lee K, Evans RM, Jaenisch R. Normal development and growth of mice carrying a targeted disruption of the a1 retinoic acid receptor gene. Proc Natl Acad Sci USA 1993; 90: 1590-4.

5. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature 1993; 366: 362-5.

6. Jüttermann R, Li E, Jaenisch R. The toxicity of 5-aza-2'-deoxycytidine to mammalian cells is mediated primarily by DNA methyltransferase rather than DNA demethylation. Proc Natl Acad Sci USA 1994; 91, 11797-11801.

7. Laird PW, Jackson-Grusby L, Fazeli A, Dickinson W, Jung E, Li E, Weinberg RA, Jaenisch R. Suppression of intestinal neoplasia by DNA hypomethylation. Cell 1995; 81:197-205.

8. Shinoda K, Lei H, Yoshii H, Nomura M, Nagano M, Shiba H, Sasaki H, Osawa Y, Ninomiya Y, Niwa O, Morohashi K, Li E. Developmental defects of the ventromedial hypothalamic nucleus and pituitary gonadotroph in the Ftz-F1 disrupted mice. Dev Dyn 1995; 204: 22-9.

9. Beard C, Li E, Jaenisch J. Loss of methylation activates Xist in somatic but not in embryonic cells. Genes Dev 1995; 9: 2325-2334.

10. Nan X, Tate P, Li E, Bird A. DNA methylation specifies chromosomal localization of MeCP2. Mole Cell Biol 1996; 16: 414-421.

11. Zhang W, Zimmer G, Chen J, Ladd D, Li E, Alt FW, Wiederrecht G, Cryan J, O'Neill EA, Seidman CE, Abbas AK, Seidman JG. T cell responses in calcineurin Aα-deficient mice. J Exp Med 1996; 183: 413-420.

12. Tucker KL, Beard C, Dausman J, Jackson-Grusby L, Laird PW, Lei H, Li E, Jaenisch R. Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. Genes Dev 1996; 10: 1008-1020.

13. Trasler JM, Trasler DG, Bestor TH, Li E, Ghibu F. Nuclear localization and expression of DNA methyltransferase in embryos is temporally correlated with postimplantation increase in DNA methylation. Dev Dyn 1996; 206: 239-247.

14. Harbers K, Müller U, Grams A, Li E, Jaenisch R, Franz T. Provirus integration into a gene encoding a ubiquitin-conjugating enzyme results in a placental defect and embryonic lethality. Proc Natl Acad Sci USA 1996; 93: 12412-7.

15. Nakamuta M, Chang B, Zsigmond E, Kobayashi K, Lei H, Ishida BY, Oka K, Li E, Chan L. Complete phenotypic characterization of apobec-1 knockout mice with a wild-type genetic background and a human apolipoprotein B transgenic background, and restoration of apolipoprotein B mRNA editing by somatic gene transfer of apobec-1. J Biol Chem 1996; 271: 25981-8.

16. Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R, Li E. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development 1996; 122: 3195 - 3205.

17. Chae T, Kwon YT, Bronson R, Dikkes P, Li E, Tsai L-H. Mice lacking p35, a neuronal specific activator of Cdk5, display cortical lamination defects, seizures, and adult lethality. Neuron 1997; 18: 29-42.

18. Simon AM, Goodneough DA, Li E, Paul DL. Female infertility in mice lacking connecxin 37. Nature 1997; 385: 525-9.

19. Oh SP, Li E. The signaling pathway mediated by type IIB activin receptor controls axial patterning and lateral asymmetry in the mouse. Genes Dev 1997; 11: 1812-1826.

20. Gong X, Li E, Klier G, Huang Q, Wu Y, Lei H, Kumar NM, Horwitz J, Gilula NB. Disruption of a3 connexin gene leads to protein degradation and cataractogenesis in mice. Cell 1997; 91: 833-843.

21. Bonventre JV, Huang Z, Taheri MR, O'Leary E, Li E, Moskowitz MA, Sapirstein, A. Cytosolic phospholipase A2 deficient mice have abnormal fertility and are protected against focal ischemic injury to the brain. Nature 1997; 390: 622-5.

22. Pradhan S, Talbot D, Sha M, Benner J, Hornstra L, Li E, Jaenisch R, Roberts RJ. Baculovirus-mediated expression and characterization of the full-length murine DNA methyltransferase. Nucl Acids Res 1997; 25: 4666-4673.

23. Wang S, Miura M, Jung Y-K, Zhu H, Li E, Yuan J. Murine caspase-11, an ICE-interacting protease is essential for the activation of ICE. Cell 1998; 92: 501-9.

24. Gu Z, Nomura M, Simpson BB, Lei H, van den Eijnden-van Raaij AJM, Donahoe PK, Li E. The type I activin receptor ActR-IB is required for egg cylinder organization and gastrulation in the Mouse. Genes Dev 1998; 12: 844-857.

25. Okano M, Xie S, Li E. Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. Nucl Acids Res 1998; 26: 2536-2540.

26. Bergeron L, Perez GI, MacDonald G, Shi L, Sun Y, Jurisicova A, Varmuza S, Latham K.E Flaws JA, Salter JC, Hara H, Moskowitz MA, Li E, Greenberg A, Tilly JL, Yuan J. Defects in regulation of apoptosis in caspase-2-deficient mice. Genes Dev 1998; 12: 1304-1314.

27. Yao TP, Oh SP, Fuchs M, Zhou ND, Ch'ng LE, Newsome D, Bronson RT, Li E, Livingston DM, Eckner R. Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. Cell 1998; 93: 361-372.

28. Chung UI, Lanske B, Lee K, Li E, Kronenberg, H. The parathyroid hormone/parathyroid hormone-related peptide receptor coordinates endochondral bone development by directly controlling chondrocyte differentiation. Proc Natl Acad Sci USA 1998; 95: 13030-5.

29. Nomura M, Li E. Roles for Smad2 in mesoderm formation, left-right patterning, and craniofacial development in mice. Nature 1998; 393: 786-790.

30. Okano M, Xie S, Li E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat Genet 1998; 19: 219-220.

31. Morohashi K, Tsuboi-Asai H, Matsushita S, Suda M, Nakashima M, Sasano H, Hataba Y, Li CL, Fukata J, Irie J, Watanabe T, Nagura H, Li E. Structural and functional abnormalities in the spleen of an mFtz-F1 gene-disrupted mouse. Blood 1999; 93: 1586-1594.

32. Gu Z, Reynolds EM, Song J, Lei H, Feijen A, Yu L, He W, MacLaughlin DT, van den Eijnden-van Raaij J, Donahoe PK, Li, E. The type I serine/threonine kinase receptor ActRIA (ALK2) is required for gastrulation of the mouse embryo. Development 1999; 126: 2551-2561.

33. Verheijen MH, Karperien M, Chung U, van Wijuen M, Heystek H, Hendriks JA, Veltmaat JM, Lanske B, Li E, Lowik CW, de Laat SW, Kronenberg HM, Defize LH. Parathyroid hormone-related peptide (PTHrP) induces parietal endoderm formation exclusively via the type I PTH/PTHrP receptor. Mech Dev 1999; 81: 151-161.

34. Song J, Oh SP, Schrewe H, Nomura M, Lei H, Okano M, Gridley T, Li E. The type II activin receptors are essential for egg cylinder growth, gastrulation and rostral head development in mice. Dev Biol 1999; 213: 157-169.

35. Xie S, Wang Z, Okano M, Nogami M, Li Y, He W, Okumura K, Li E. Cloning, expression, and chromosome locations of the human DNMT3 gene family. Gene 1999; 236: 87-95.

36. Donohoe MA, Zhang X-L, McGinnis L, Biggers J, Li E, Shi Y. Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality. Mol Cell Biol 1999; 19: 7237-7244.

37. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 1999; 99: 247-257.

38. Okano M, Takebayashi S, Okumura K, Li E. Assignment of cytosine-5 DNA methyltransferases Dnmt3a and Dnmt3b to mouse chromosome bands 12A2-A3 and 2H1 by in situ hybridization. Cytogenet Cell Genet 1999; 86: 333-334.

39. Li YP, Chen W, Liang Y, Li E, Stashenko P. OC-116Kda-deficient mice exibit severe osteopetrosis due to loss of osteoclast-mediated extracellular acidification. Nat Genet 1999; 23: 452-456.

40. Nakagawa T, Zhu H, Morishima N, Li E, Xu J, Yankner BA, Yuan J. Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-beta. Nature 2000; 403:98-103.

41. Oh, SP., Seki, T., Goss, KA., Imamura, T., Yi, Y., Donahoe, PK., Li, L., Miyazono, K., ten Dijke, P., Kim, S., and Li, E. Activin Receptor-Like Kinase 1 (ALK1) modulates TGF-b1 signaling in the regulation of angiogenesis. Proc Natl Acad Sci USA 2000 (in press).

42. Sado, T., Fenner, M. H., Tan, S.-S., Tam, P., Shioda, T., and **Li, E.** (2000). X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. *Dev. Biol.* 15: 294-303.

43. Kim, S.K., Hebrok, M., Li, E., Oh, S.P., Schrewe, H., Harmon, E.B., Lee, J.S., and Melton, D.A. (2000). Activin receptor patterning of foregut organogenesis. Genes Dev 14:1866-1871.

44. Pannell D, Osborne CS, Yao S, Sukonnik T, Pasceri P, Karaiskakis A, Okano M, Li E, Lipshitz HD, and Ellis J (2000) Retrovirus vector silencing is de novo methylase independent and marked by a repressive histone code. EMBO J. 19:5884-5894.

45. Sado, T., Wang, Z., Sasaki, H., and Li, E. (2001). Regulation of Imprinted X-Inactivation in Mice by Tsix. Development 128: 1275-1286.

46. Ferguson, C.A., Tucker, A.S., Heikinheimo, K., Nomura, M., Oh, P., Li, E., Sharpe, PT. (2001) The role of effectors of the activin signalling pathway, activin receptors IIA and IIB, and Smad2, in patterning of tooth. Development 128:4605-4613.

47. Ko, J., Humbert, S., Bronson, R.T., Takahashi, S., Kulkarni, A.B., Li, E., Tsai, L.H. (2001) p35 and p39 are essential for cyclin-dependent kinase 5 function during neurodevelopment. J Neurosci 21:6758-6771.

48. Jiang P, Song J, Gu G, Slonimsky E, Li E, Rosenthal N. (2002). Targeted deletion of the MLC1f/3f downstream enhancer results in precocious MLC expression and mesoderm ablation. Dev Biol 243:281-293.

49. Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, Laird PW, Jones PA. (2002). Cooperativity between DNA Methyltransferases in the Maintenance Methylation of Repetitive Elements. Mol Cell Biol 22(2):480-491.

50. Hata, H., Okano, M., Lei, H., and Li, E. (2002) Dnmt3L cooperates with the Dnmt3 family de novo DNA methyltransferases to establish maternal imprinting in mice. Development 129: 1983-1993.

51. Sakuma R, Ohnishi Yi Y, Meno C, Fujii H, Juan H, Takeuchi J, Ogura T, Li E, Miyazono K, Hamada H. (2002) Inhibition of Nodal signalling by Lefty mediated through interaction with common receptors and efficient diffusion. Genes Cells 7:401-412.

52. Miura K, Kishino T, Li E, Webber H, Dikkes P, Holmes GL, Wagstaff J. (2002) Neurobehavioral and electroencephalographic abnormalities in ube3a maternal-deficient mice. Neurobiol Dis 9:149-159.

53. Dodge, J. Ramsahoye, B.H., Wo, Z.G, Okano, M., and Li, E. (2002) *De novo* methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. Gene 289:41-48.

54. Oh. S.P., and Li, E. (2002). Gene-dosage sensitive genetic interactions between inversus viserum (*iv*), nodal, and activin type IIB receptor (*ActRIIB*) genes in asymmetrical patterning of the visceral organs along the left-right axis. Dev. Dyn. 224:279-290.

55. Chen T, Ueda Y, Xie S, Li E. (2002). A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. J Biol Chem. 277(41):38746-54.

56. Oh, SP, Yeo, CY, Lee, Y, Schrewe, H, Whitman, M, and Li, E (2002). Activin type IIA and type IIB receptors mediate Gdf11 signaling in axial vertebral patterning. Genes Dev. 16: 2749-2754.

57. Sado T, Li E, Sasaki H. Effect of TSIX disruption on XIST expression in male ES cells. Cytogenet Genome Res. 2002;99:115-118.

58. Trasler J, Deng L, Melnyk S, Pogribny I, Hiou-Tim F, Sibani S, Oakes C, Li E, James SJ, Rozen R. Impact of Dnmt1 deficiency, with and without low folate diets, on tumor numbers and DNA methylation in Min mice. Carcinogenesis. 2003 Jan;24(1):39-45.

59. Welte T, Zhang SS, Wang T, Zhang Z, Hesslein DG, Yin Z, Kano A, Iwamoto Y, Li E, Craft JE, Bothwell AL, Fikrig E, Koni PA, Flavell RA, Fu XY(2003). STAT3 deletion during hematopoiesis causes Crohn's disease-like pathogenesis and lethality: a critical role of STAT3 in innate immunity. Proc Natl Acad Sci U S A. 2003 Feb 18;100(4):1879-84.

60. Wei Chen, Yuqiong Liang, Wenjie Deng, Ken Shimizu, Amir M. Ashique, En Li, and Yi-Ping Li (2003). The zinc-finger protein CNBP is required for forebrain formation in the mouse. Development 130: 1367-1379.

61. Joost Gribnau, Konrad Hochedlinger, Ken Hata, En Li, and Rudolf Jaenisch (2003). Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. Genes Dev 17:759-773.

62. Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T, Peters AH. (2003) Suv39h-mediated histone h3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol. 13:1192-1200.

63. Genomic organization and promoter analysis of the Dnmt3b gene. (2003) Ishida C, Ura K, Hirao A, Sasaki H, Toyoda A, Sakaki Y, Niwa H, Li E, Kaneda Y. Gene. 310:151-9.

64. Taiping Chen, Yoshihide Ueda, Jonathan E. Dodge, Zhenjuan Wang, and En Li (2003) Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. Mole. Cell Biol., 23:5594-5605.

65. Jacoby JJ, Kalinowski A, Liu MG, Zhang SS, Gao Q, Chai GX, Ji L, Iwamoto Y, Li E, Schneider M, Russell KS, Fu XY. Cardiomyocyte-restricted knockout of STAT3 results in higher sensitivity to inflammation, cardiac fibrosis, and heart failure with advanced age. Proc Natl Acad Sci U S A. 2003. 100:12929-34.

66. Kelly TL, Li E, Trasler JM. 5-aza-2'-deoxycytidine induces alterations in murine spermatogenesis and pregnancy outcome. J Androl. 2003. 24:822-830.

67. Mund C, Musch T, Strodicke M, Assmann B, Li E, Lyko F. Comparative analysis of DNA methylation patterns in transgenic Drosophila overexpressing mouse DNA methyltransferases. Biochem J. 2004. 378:763-768.

68. Sado T, Okano M, Li E, Sasaki H. De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. Development. 2004. 131:975-982.

69. Dodge JE, Kang YK, Beppu H, Lei H, Li E. Histone H3-K9 methyltransferase ESET is essential for early development. Mol Cell Biol. 2004. 24:2478-86.

70. Kaneda M, Okano M, Hata K, Sado T, Tsujimoto N, Li E, Sasaki H. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature. 2004. 429(6994):900-3.

71. Beppu H, Ichinose F, Kawai N, Jones RC, Yu PB, Zapol WM, Miyazono K, Li E, Bloch KD. BMPR-II heterozygous mice have mild pulmonary hypertension and an impaired pulmonary vascular remodeling response to prolonged hypoxia. Am J Physiol Lung Cell Mol Physiol. 2004 Jul 30

72. de Sousa Lopes SM, Roelen BA, Monteiro RM, Emmens R, Lin HY, Li E, Lawson KA, Mummery CL. BMP signaling mediated by ALK2 in the visceral endoderm is necessary for the generation of primordial germ cells in the mouse embryo. Genes Dev. 2004 Aug 1;18(15):1838-49.

73. Hattori N, Abe T, Hattori N, Suzuki M, Matsuyama T, Yoshida S, Li E, Shiota K. Preference of DNA methyltransferases for CpG islands in mouse embryonic stem cells. Genome Res. 2004. 14(9):1733-40.

74. Chen T, Tsujimoto N, Li E. The PWWP Domain of Dnmt3a and Dnmt3b Is Required for Directing DNA Methylation to the Major Satellite Repeats at Pericentric Heterochromatin. Mol Cell Biol. 2004. 24:9048-58.

75. Fang J, Chen T, Chadwick B, Li E, Zhang Y. Ring1b-mediated H2A ubiquitination associates with inactive X chromosomes and is involved in initiation of X inactivation. J Biol Chem. 2004 Dec 17;279(51):52812-5.

76. Beppu H, Ichinose F, Kawai N, Jones RC, Yu PB, Zapol WM, Miyazono K, Li E, Bloch KD. BMPR-II heterozygous mice have mild pulmonary hypertension and an impaired pulmonary vascular remodeling response to prolonged hypoxia. Am J Physiol Lung Cell Mol Physiol. 2004 Dec;287(6):L1241-7.

77. Park S, Lee YJ, Lee HJ, Seki T, Hong KH, Park J, Beppu H, Lim IK, Yoon JW, Li E, Kim SJ, Oh SP. B-cell translocation gene 2 (Btg2) regulates vertebral patterning by modulating bone morphogenetic protein/smad signaling. Mol Cell Biol. 2004 Dec;24(23):10256-62.

78. Feng J, Chang H, Li E, Fan G. Dynamic expression of de novo DNA methyltransferases Dnmt3a and Dnmt3b in the central nervous system. J Neurosci Res. 2005 Mar 15;79(6):734-46.

79. Beppu H, Lei H, Bloch KD, Li E. Generation of a floxed allele of the mouse BMP type II receptor gene. Genesis. 2005 Feb 25;41(3):133-137

80. Yu PB, Beppu H, Kawai N, Li E, Bloch KD. Bone morphogenetic protein (BMP) type II receptor deletion reveals BMP ligand-specific gain of signaling in pulmonary artery smooth muscle cells. J Biol Chem. 2005. 280:24443-50.

81. Hopfer U, Fukai N, Hopfer H, Wolf G, Joyce N, Li E, Olsen BR. Targeted disruption of Col8a1 and Col8a2 genes in mice leads to anterior segment abnormalities in the eye. FASEB J. 2005. 19:1232-44.

82. Rodic N, Oka M, Hamazaki T, Murawski MR, Jorgensen M, Maatouk DM, Resnick JL, Li E, Terada N. DNA methylation is required for silencing of Ant4, an adenine nucleotide translocase selectively expressed in mouse embryonic stem cells and germ cells. Stem Cells. 2005 Jul 28; [Epub ahead of print]

Proceedings of Meetings

1. Li E, Beard C, Foster A, Bestor TH, Jaenisch R. DNA methylation, genomic imprinting, and mammalian development. Proceedings of the Cold Spring Harbor Symposium on Quantitative Biology; 1993; 58: 297-305.
2. Muragaki Y, Timmons S, Griffith CM, Oh SP, Li E, Fukai N, Fadel B, Quertermous T,

Olsen B. Tissue-specific expression of three alternative variants of *Col18*a1 and their localization in basement membrane zones. Proceedings of the 7th International Symposium on Basement Membranes, Bethesda, 1995.
3. Okano M, Li E. (2002). Genetic analyses of DNA methyltransferase genes in mouse model system. J Nutr. 132:2462S-5S.

Reviews, Book Chapters, and Editorials

1. Jaenisch R, Beard C, Li, E. DNA methylation and mammalian development. In: Ohlsson R, Hall K, Ritzen M, editors. Genomic imprinting: Causes and consequences. Cambridge, UK: Cambridge University Press; 1995. p. 118.
2. Li, E. Role of DNA methylation in mammalian development. In: Reik W, Sorani A, editors. Genomic Imprinting - Frontiers in Molecular Biology. Volume 18. Oxford, UK: Oxford University Press; 1997. p. 1-20.
3. Li, E. The mojo of methylation [News & Views]. Nat Genet 1999; 23: 5-6.
4. Li, E and Jaenisch, R. DNA methylation and methyltransferases. In: Ehrlich M, editor. DNA alterations in cancer: Genetic and epigenetic changes. PP. 351-365. Natick: BioTechniques Books; 2000.
5. Li E. Chromatin modification and epigenetic reprogramming in mammalian development. 2002. Nature Rev. Genet. 3, 662-673.
6. Chen T, Li E. Structure and function of eukaryotic DNA methyltransferases. 2004. Curr Top Dev Biol. 60:55-89.

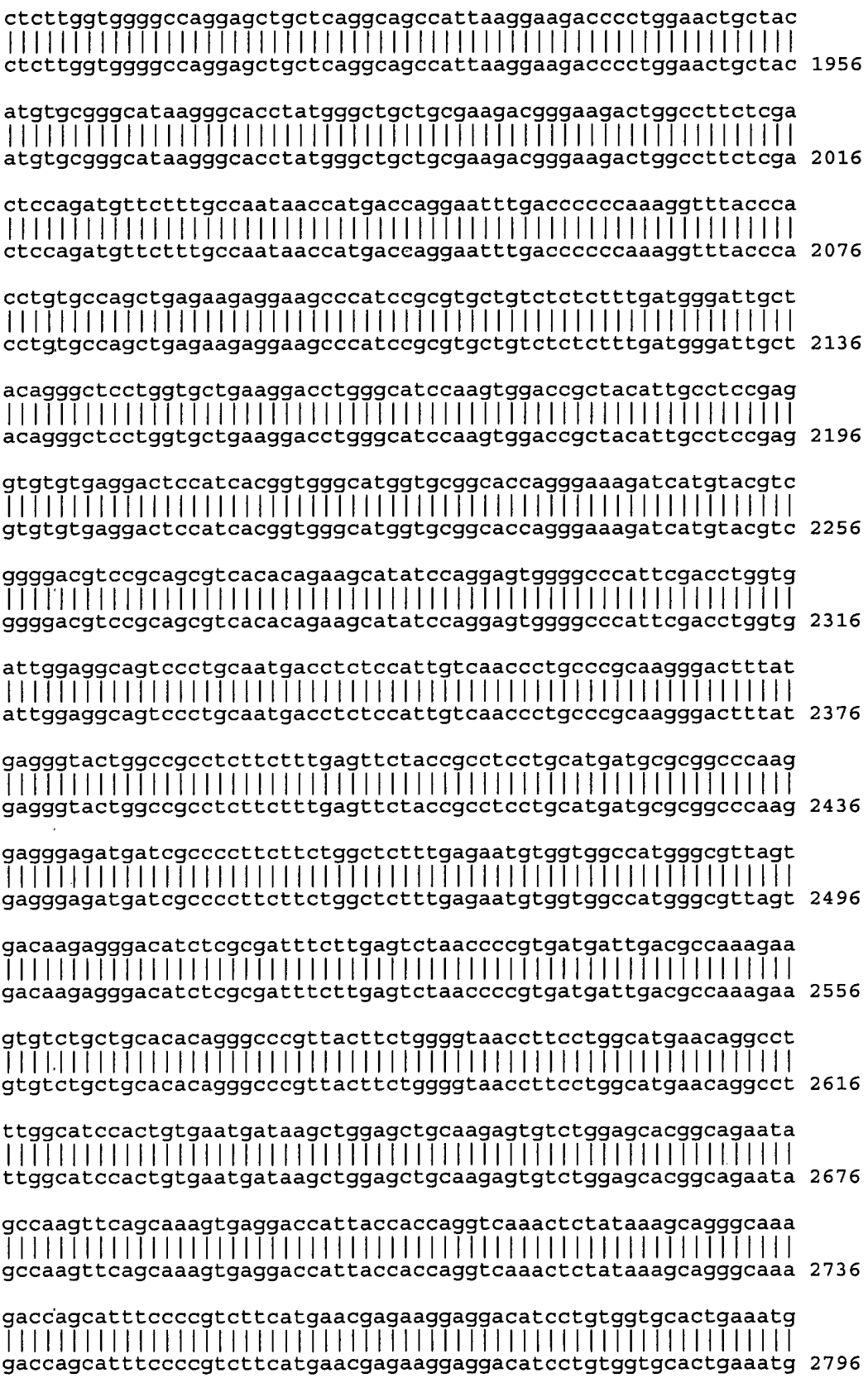
EXHIBIT B

Alignment spanning the coding region of mouse Dnmt3a sequence from ATCC Deposit No. 209933 (top) and currently amended SEQ ID NO:1 (bottom)[1]

```
atgccctccagcggccccggggacaccagcagctcctctctggagcgggaggatgatcga
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
atgccctccagcggccccggggacaccagcagctcctctctggagcgggaggatgatcga 276

aaggaaggagaggaacaggaggagaaccgtggcaaggaagagcgccaggagcccagcgcc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aaggaaggagaggaacaggaggagaaccgtggcaaggaagagcgccaggagcccagcgcc 336

acggcccggaaggtggggaggcctggccggaagcgcaagcacccaccggtggaaagcagt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
acggcccggaaggtggggaggcctggccggaagcgcaagcacccaccggtggaaagcagt 396

gacacccccaaggacccagcagtgaccaccaagtctcagcccatggcccaggactctggc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gacacccccaaggacccagcagtgaccaccaagtctcagcccatggcccaggactctggc 456

ccctcagatctgctacccaatggagacttggagaagcggagtgaaccccaacctgaggag
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ccctcagatctgctacccaatggagacttggagaagcggagtgaaccccaacctgaggag 516

gggagcccagctgcagggcagaagggtggggccccagctgaaggagagggaactgagacc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gggagcccagctgcagggcagaagggtggggccccagctgaaggagagggaactgagacc 576

ccaccagaagcctccagagctgtggagaatggctgctgtgtgaccaaggaaggccgtgga
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ccaccagaagcctccagagctgtggagaatggctgctgtgtgaccaaggaaggccgtgga 636

gcctctgcaggagagggcaaagaacagaagcagaccaacatcgaatccatgaaaatggag
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gcctctgcaggagagggcaaagaacagaagcagaccaacatcgaatccatgaaaatggag 696

ggctcccggggccgactgcgaggtggcttgggctgggagtccagcctccgtcagcgaccc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ggctcccggggccgactgcgaggtggcttgggctgggagtccagcctccgtcagcgaccc 756

atgccaagactcaccttccaggcaggggacccctactacatcagcaaacggaaacgggat
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
atgccaagactcaccttccaggcaggggacccctactacatcagcaaacggaaacgggat 816

gagtggctggcacgttggaaaagggaggctgagaagaaagccaaggtaattgcagtaatg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gagtggctggcacgttggaaaagggaggctgagaagaaagccaaggtaattgcagtaatg 876

aatgctgtggaagagaaccaggcctctggagagtctcagaaggtggaggaggccagccct
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aatgctgtggaagagaaccaggcctctggagagtctcagaaggtggaggaggccagccct 936

cctgctgtgcagcagcccacggaccctgcttctccgactgtggccaccacccctgagcca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cctgctgtgcagcagcccacggaccctgcttctccgactgtggccaccacccctgagcca 996
```

[1] Bolded nucleotides indicate nucleotides that were amended on July 23, 2001.

```
gtaggaggggatgctggggacaagaatgctaccaaagcagccgacgatgagcctgagtat
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gtaggaggggatgctggggacaagaatgctaccaaagcagccgacgatgagcctgagtat 1056

gaggatggccggggctttggcattggagagctggtgtgggggaaacttcggggcttctcc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gaggatggccggggctttggcattggagagctggtgtgggggaaacttcggggcttctcc 1116

tggtggccaggccgaattgtgtcttggtggatgacaggccggagccgagcagctgaaggc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
tggtggccaggccgaattgtgtcttggtggatgacaggccggagccgagcagctgaaggc 1176

actcgctgggtcatgtggttcggagatggcaagttctcagtggtgtgtgtggagaagctc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
actcgctgggtcatgtggttcggagatggcaagttctcagtggtgtgtgtggagaagctc 1236

atgccgctgagctccttctgcagtgcattccaccaggccacctacaacaagcagcccatg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
atgccgctgagctccttctgcagtgcattccaccaggccacctacaacaagcagcccatg 1296

taccgcaaagccatctacgaagtcctccaggtggccagcagccgtgccgggaagctgttt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
taccgcaaagccatctacgaagtcctccaggtggccagcagccgtgccgggaagctgttt 1356

ccagcttgccatgacagtgatgaaagtgacagtggcaaggctgtggaagtgcagaacaag
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ccagcttgccatgacagtgatgaaagtgacagtggcaaggctgtggaagtgcagaacaag 1416

cagatgattgaatgggccctcggtggcttccagccctcgggtcctaagggcctggagcca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cagatgattgaatgggccctcggtggcttccagccctcgggtcctaagggcctggagcca 1476

ccagaagaagagaagaatccttacaaggaagtttacaccgacatgtgggtggagcctgaa
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ccagaagaagagaagaatccttacaaggaagtttacaccgacatgtgggtggagcctgaa 1536

gcagctgcttacgccccacccccaccagccaagaaacccagaaagagcacaacagagaaa
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gcagctgcttacgccccacccccaccagccaagaaacccagaaagagcacaacagagaaa 1596

cctaaggtcaaggagatcattgatgagcgcacaagggagcggctggtgtatgaggtgcgc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cctaaggtcaaggagatcattgatgagcgcacaagggagcggctggtgtatgaggtgcgc 1656

cagaagtgcagaaacatcgaggacatttgtatctcatgtgggagcctcaatgtcaccctg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cagaagtgcagaaacatcgaggacatttgtatctcatgtgggagcctcaatgtcaccctg 1716

gagcacccactcttcattggaggcatgtgccagaactgtaagaactgcttcttggagtgt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gagcacccactcttcattggaggcatgtgccagaactgtaagaactgcttcttggagtgt 1776

gcttaccagtatgacgacgatgggtaccagtcctattgcaccatctgctgtgggggcgt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gcttaccagtatgacgacgatgggtaccagtcctattgcaccatctgctgtgggggcgt 1836

gaagtgctcatgtgtgggaacaacaactgctgcaggtgcttttgtgtcgagtgtgtggat
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gaagtgctcatgtgtgggaacaacaactgctgcaggtgcttttgtgtcgagtgtgtggat 1896
```


ctcttggtggggccaggagctgctcaggcagccattaaggaagacccctggaactgctac
ctcttggtggggccaggagctgctcaggcagccattaaggaagacccctggaactgctac 1956
atgtgcgggcataagggcacctatgggctgctgcgaagacgggaagactggccttctcga
atgtgcgggcataagggcacctatgggctgctgcgaagacgggaagactggccttctcga 2016
ctccagatgttctttgccaataaccatgaccaggaatttgacccccaaaggtttaccca
ctccagatgttctttgccaataaccatgaccaggaatttgacccccaaaggtttaccca 2076
cctgtgccagctgagaagaggaagcccatccgcgtgctgtctctctttgatgggattgct
cctgtgccagctgagaagaggaagcccatccgcgtgctgtctctctttgatgggattgct 2136
acagggctcctggtgctgaaggacctgggcatccaagtggaccgctacattgcctccgag
acagggctcctggtgctgaaggacctgggcatccaagtggaccgctacattgcctccgag 2196
gtgtgtgaggactccatcacggtgggcatggtgcggcaccagggaaagatcatgtacgtc
gtgtgtgaggactccatcacggtgggcatggtgcggcaccagggaaagatcatgtacgtc 2256
ggggacgtccgcagcgtcacacagaagcatatccaggagtggggcccattcgacctggtg
ggggacgtccgcagcgtcacacagaagcatatccaggagtggggcccattcgacctggtg 2316
attggaggcagtccctgcaatgacctctccattgtcaaccctgcccgcaagggactttat
attggaggcagtccctgcaatgacctctccattgtcaaccctgcccgcaagggactttat 2376
gagggtactggccgcctcttctttgagttctaccgcctcctgcatgatgcgcggcccaag
gagggtactggccgcctcttctttgagttctaccgcctcctgcatgatgcgcggcccaag 2436
gagggagatgatcgccccttcttctggctctttgagaatgtggtggccatgggcgttagt
gagggagatgatcgccccttcttctggctctttgagaatgtggtggccatgggcgttagt 2496
gacaagagggacatctcgcgatttcttgagtctaaccccgtgatgattgacgccaaagaa
gacaagagggacatctcgcgatttcttgagtctaaccccgtgatgattgacgccaaagaa 2556
gtgtctgctgcacacagggcccgttacttctggggtaaccttcctggcatgaacaggcct
gtgtctgctgcacacagggcccgttacttctggggtaaccttcctggcatgaacaggcct 2616
ttggcatccactgtgaatgataagctggagctgcaagagtgtctggagcacggcagaata
ttggcatccactgtgaatgataagctggagctgcaagagtgtctggagcacggcagaata 2676
gccaagttcagcaaagtgaggaccattaccaccaggtcaaactctataaagcagggcaaa
gccaagttcagcaaagtgaggaccattaccaccaggtcaaactctataaagcagggcaaa 2736
gaccagcatttccccgtcttcatgaacgagaaggaggacatcctgtggtgcactgaaatg
gaccagcatttccccgtcttcatgaacgagaaggaggacatcctgtggtgcactgaaatg 2796

```
gaaagggtgtttggcttccccgtccactacacagacgtctccaacatgagccgcttggcg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gaaagggtgtttggcttccccgtccactacacagacgtctccaacatgagccgcttggcg 2856

aggcagagactgctgggccgatcgtggagcgtgccggtcatccgccacctcttcgctccg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aggcagagactgctgggccgatcgtggagcgtgccggtcatccgccacctcttcgctccg 2916

ctgaaggaatattttgcttgtgtgtaagggacatgggggcaaactgaagtagtgatgata
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ctgaaggaatattttgcttgtgtgtaagggacatgggggcaaactgaagtagtgatgata 2976

aaaaagttaaacaaacaaacaaacaccaagaacgagaggacggagaaaagttcagcaccc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aaaaagttaaacaaacaaacaaacaccaagaacgagaggacggagaaaagttcagcaccc 3037

agaagagaaaaaggaatttaaagcaaaccacagaggaggaaaacgccggagggcttggcc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
agaagagaaaaaggaatttaaagcaaaccacagaggaggaaaacgccggagggcttggcc 3098

ttgcaaaagggttggacatcatctcctgagttttcaatgttaaccttcagtcctatctaa
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ttgcaaaagggttggacatcatctcctgagttttcaatgttaaccttcagtcctatctaa 3158

aaagcaaaataggcccctcccctttcttcccctccggtcctaggaggcgaacttttgttt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aaagcaaaataggcccctcccctttcttcccctccggtcctaggaggcgaacttttgttt 3218

tctactctttttcagaggggttttctgtttgtttgggtttttgtttcttgctgtgactga
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
tctactctttttcagaggggttttctgtttgtttgggtttttgtttcttgctgtgactga 3278

aacaagagagttattgcagcaaaatcagtaacaacaaaaagtagaaatgccttggagagg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aacaagagagttattgcagcaaaatcagtaacaacaaaaagtagaaatgccttggagagg 3338

aaagggagagagggaaaattctataaaaacttaaaatattggtttttttttttttccctt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aaagggagagagggaaaattctataaaaacttaaaatattggtttttttttttttccctt 3398

ttctatatatctctttggttgtctctagcctgatcagataggagcacaaacaggaagaga
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ttctatatatctctttggttgtctctagcctgatcagataggagcacaaacaggaagaga 3458

atagagaccctcggaggcagagtctcctctcccacccccgagcagtctcaacagcacca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
atagagaccctcggaggcagagtctcctctcccacccccgagcagtctcaacagcacca 3518

ttcctggtcatgcaaaacagaacccaactagcagcagggcgctgagagaacaccacacca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ttcctggtcatgcaaaacagaacccaactagcagcagggcgctgagagaacaccacacca 3578

gacactttctacagtatttcaggtgcctaccacacaggaaaccttgaagaaaaccagttt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gacactttctacagtatttcaggtgcctaccacacaggaaaccttgaagaaaaccagttt 3638

ctagaagccgctgttacctcttgtttacagtt
|||||||||||||||||||||||||||||||
ctagaagccgctgttacctcttgtttacagtt 3670
```

EXHIBIT C

Alignment spanning the coding region of mouse Dnmt3b from ATCC Deposit No. 209934 (top) and currently amended SEQ ID NO:2 (bottom)[2]

```
caggaaacaatgaagggagacagcagacatctgaatgaagaagagggtgccagcgggtat
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
caggaaacaatgaagggagacagcagacatctgaatgaagaagagggtgccagcgggtat 319

gaggagtgcattatcgttaatgggaacttcagtgaccagtcctcagacacgaaggatgct
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gaggagtgcattatcgttaatgggaacttcagtgaccagtcctcagacacgaaggatgct 379

ccctcacccccagtcttggaggcaatctgcacagagccagtctgcacaccagagaccaga
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ccctcacccccagtcttggaggcaatctgcacagagccagtctgcacaccagagaccaga 439

ggccgcaggtcaagctcccggctgtctaagagggaggtctccagccttctgaattacacg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ggccgcaggtcaagctcccggctgtctaagagggaggtctccagccttctgaattacacg 499

caggacatgacaggagatggagacagagatgatgaagtagatgatgggaatggctctgat
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
caggacatgacaggagatggagacagagatgatgaagtagatgatgggaatggctctgat 559

attctaatgccaaagctcacccgtgagaccaaggacaccaggacgcgctctgaaagcccg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
attctaatgccaaagctcacccgtgagaccaaggacaccaggacgcgctctgaaagcccg 619

gctgtccgaacccgacatagcaatgggacctccagcttggagaggcaaagagcctccccc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gctgtccgaacccgacatagcaatgggacctccagcttggagaggcaaagagcctccccc 679

agaatcacccgaggtcggcagggccgccaccatgtgcaggagtaccctgtggagtttccg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
agaatcacccgaggtcggcagggccgccaccatgtgcaggagtaccctgtggagtttccg 739

gctaccaggtctcggagacgtcgagcatcgtcttcagcaagcacgccatggtcatccccct
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gctaccaggtctcggagacgtcgagcatcgtcttcagcaagcacgccatggtcatccccct 799

gccagcgtcgacttcatggaagaagtgacacctaagagcgtcagtaccccatcagttgac
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gccagcgtcgacttcatggaagaagtgacacctaagagcgtcagtaccccatcagttgac 859

ttgagccaggatggagatcaggagggtatggataccacacaggtggatgcagagagcaga
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ttgagccaggatggagatcaggagggtatggataccacacaggtggatgcagagagcaga 919

gatggagacagcacagagtatcaggatgataaagagtttggaataggtgacctcgtgtgg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gatggagacagcacagagtatcaggatgataaagagtttggaataggtgacctcgtgtgg 979



ggaaagatcaagggcttctcctggtggcctgccatggtggtgtcctggaaagccacctcc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ggaaagatcaagggcttctcctggtggcctgccatggtggtgtcctggaaagccacctcc 1039
```

[2] Bolded nucleotides indicate nucleotides that were amended on July 23, 2001.

```
aagcgacaggccatgcccggaatgcgctgggtacagtggtttggtgatggcaagttttct
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aagcgacaggccatgcccggaatgcgctgggtacagtggtttggtgatggcaagttttct 1099

gagatctctgctgacaaactggtggctctggggctgttcagccagcactttaatctggct
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gagatctctgctgacaaactggtggctctggggctgttcagccagcactttaatctggct 1159

accttcaataagctggtttcttataggaaggccatgtaccacactctggagaaagccagg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accttcaataagctggtttcttataggaaggccatgtaccacactctggagaaagccagg 1219

gttcgagctggcaagaccttctccagcagtcctggagagtcactggaggaccagctgaag
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gttcgagctggcaagaccttctccagcagtcctggagagtcactggaggaccagctgaag 1279

cccatgctggagtgggcccacggtggcttcaagcctactgggatcgagggcctcaaaccc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cccatgctggagtgggcccacggtggcttcaagcctactgggatcgagggcctcaaaccc 1339

aacaagaagcaaccagtggttaataagtcgaaggtgcgtcgttcagacagtaggaactta
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aacaagaagcaaccagtggttaataagtcgaaggtgcgtcgttcagacagtaggaactta 1399

gaacccaggagacgcgagaacaaaagtcgaagacgcacaaccaatgactctgctgcttct
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gaacccaggagacgcgagaacaaaagtcgaagacgcacaaccaatgactctgctgcttct 1459

gagtccccccccacccaagcgcctcaagacaaatagctatggcgggaaggaccgaggggag
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gagtccccccccacccaagcgcctcaagacaaatagctatggcgggaaggaccgaggggag 1519

gatgaggagagccgagaacggatggcttctgaagtcaccaacaacaagggcaatctggaa
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gatgaggagagccgagaacggatggcttctgaagtcaccaacaacaagggcaatctggaa 1579

gaccgctgtttgtcctgtggaaagaagaaccctgtgtccttccacccccctctttgagggt
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gaccgctgtttgtcctgtggaaagaagaaccctgtgtccttccacccccctctttgagggt 1639

gggctctgtcagagttgccgggatcgcttcctagagctcttctacatgtatgatgaggac
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gggctctgtcagagttgccgggatcgcttcctagagctcttctacatgtatgatgaggac 1699

ggctatcagtcctactgcaccgtgtgctgtgagggccgtgaactgctgctgtgcagtaac
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ggctatcagtcctactgcaccgtgtgctgtgagggccgtgaactgctgctgtgcagtaac 1759

acaagctgctgcagatgcttctgtgtggagtgtctggaggtgctggtgggcgcaggcaca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
acaagctgctgcagatgcttctgtgtggagtgtctggaggtgctggtgggcgcaggcaca 1819

gctgaggatgccaagctgcaggaaccctggagctgctatatgtgcctccctcagcgctgc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gctgaggatgccaagctgcaggaaccctggagctgctatatgtgcctccctcagcgctgc 1879


catggggtcctccgacgcaggaaagattggaacatgcgcctgcaagacttcttcactact
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
catggggtcctccgacgcaggaaagattggaacatgcgcctgcaagacttcttcactact 1939
```

```
gatcctgacctggaagaatttgagccacccaagttgtacccagcaattcctgcagccaaa
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gatcctgacctggaagaatttgagccacccaagttgtacccagcaattcctgcagccaaa 1999

aggaggcccattagagtcctgtctctgtttgatggaattgcaacggggtacttggtgctc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aggaggcccattagagtcctgtctctgtttgatggaattgcaacggggtacttggtgctc 2059

aaggagttgggtattaaagtggaaaagtacattgcctccgaagtctgtgcagagtccatc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aaggagttgggtattaaagtggaaaagtacattgcctccgaagtctgtgcagagtccatc 2119

gctgtgggaactgttaagcatgaaggccagatcaaatatgtcaatgacgtccggaaaatc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gctgtgggaactgttaagcatgaaggccagatcaaatatgtcaatgacgtccggaaaatc 2179

accaagaaaaatattgaagagtggggcccgttcgacttggtgattggtggaagcccatgc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
accaagaaaaatattgaagagtggggcccgttcgacttggtgattggtggaagcccatgc 2239

aatgatctctctaacgtcaatcctgcccgcaaaggtttatatgagggcacaggaaggctc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aatgatctctctaacgtcaatcctgcccgcaaaggtttatatgagggcacaggaaggctc 2299

ttcttcgagttttaccacttgctgaattatacccgccccaaggagggcgacaaccgtcca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ttcttcgagttttaccacttgctgaattatacccgccccaaggagggcgacaaccgtcca 2359

ttcttctggatgttcgagaatgttgtggccatgaaagtgaatgacaagaaagacatctca
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ttcttctggatgttcgagaatgttgtggccatgaaagtgaatgacaagaaagacatctca 2419

agattcctggcatgtaacccagtgatgatcgatgccatcaaggtgtctgctgctcacagg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
agattcctggcatgtaacccagtgatgatcgatgccatcaaggtgtctgctgctcacagg 2479

gcccggtacttctggggtaacctacccggaatgaacaggcccgtgatggcttcaaagaat
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gcccggtacttctggggtaacctacccggaatgaacaggcccgtgatggcttcaaagaat 2539

gataagctcgagctgcaggactgcctggagttcagtaggacagcaaagttaaagaaagtg
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gataagctcgagctgcaggactgcctggagttcagtaggacagcaaagttaaagaaagtg 2599

cagacaataaccaccaagtcgaactccatcagacagggcaaaaaccagcttttccctgta
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cagacaataaccaccaagtcgaactccatcagacagggcaaaaaccagcttttccctgta 2659

gtcatgaatggcaaggacgacgttttgtggtgcactgagctcgaaaggatcttcggcttc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
gtcatgaatggcaaggacgacgttttgtggtgcactgagctcgaaaggatcttcggcttc 2719

cctgctcactacacggacgtgtccaacatgggccgcggcgcccgtcagaagctgctgggc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
cctgctcactacacggacgtgtccaacatgggccgcggcgcccgtcagaagctgctgggc 2779

aggtcctggagtgtaccggtcatcagacacctgtttgccccttgaaggactactttgcc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
aggtcctggagtgtaccggtcatcagacacctgtttgccccttgaaggactactttgcc 2839
```

```
tgtgaatagttctacccaggactggggagctctcggtcagagccagtgcccagagtc
||||||||||||||||||||||||||||||||||||||||||||||||||||||||
tgtgaatagttctacccaggactggggagctctcggtcagagccagtgcccagagtc 2896
```

LETTER

Estimation of Errors in "Raw" DNA Sequences: A Validation Study

Peter Richterich[1]

Genome Therapeutics Corp., Waltham, Massachusetts 02154 USA

As DNA sequencing is performed more and more in a mass-production-like manner, efficient quality control measures become increasingly important for process control, but so also does the ability to compare different methods and projects. One of the fundamental quality measures in sequencing projects is the position-specific error probability at all bases in each individual sequence. Accurate prediction of base-specific error rates from "raw" sequence data would allow immediate quality control as well as benchmarking different methods and projects while avoiding the inefficiencies and time delays associated with resequencing and assessments after "finishing" a sequence. The program PHRED provides base-specific quality scores that are logarythmically related to error probabilities. This study assessed the accuracy of PHRED's error-rate prediction by analyzing sequencing projects from six different large-scale sequencing laboratories. All projects used four-color fluorescent sequencing, but the sequencing methods used varied widely between the different projects. The results indicate that the error-rate predictions such as those given by PHRED can be highly accurate for a large variety of different sequencing methods as well as over a wide range of sequence quality.

In DNA sequencing, knowledge about the accuracy of sequences can be very valuable. For example, different large-scale sequencing projects may produce sequences at similar rates and costs but with significantly different error rates in the final sequence. One major determinant in the final error rate is the accuracy of the "raw" sequence. Knowledge about the frequency and location of errors in the raw sequence data can help to direct "polishing" efforts to the places where additional effort is needed; it also enables the comparison between different sequencing projects without requiring that the same region be sequenced in each project.

Another area where estimates about sequence error rates would be beneficial is technology development. Accurate error estimates at each base would enable "quality benchmarking" between different methods, thus enabling researchers to choose the method that fills their needs for accuracy and throughput best.

Several groups have developed mathematical models to predict the error probability at any given position in raw sequences. Lawrence and Solovyev used linear discriminant analysis to calculate separate probability estimates for insertions, deletions, and mismatches (Lawrence and Solovyev 1994). Ewing and Green (1998) developed the program PHRED, which calculates a quality score at each base. This quality score q is logarithmically linked to the error probability p: $q = -10 \times \log_{10}(p)$ (for a discussion of how quality scores are calculated and what the limitations are, see Ewing et al. (1998). When used in combination with sequence assembly and finishing programs that utilize these error estimates, reliable error probabilities promise to increase the accuracy of consensus sequences and to reduce the efforts required in the finishing phase of sequencing projects (Churchill and Waterman 1992; Bonfield and Staden 1995).

To examine the accuracy of probability estimates made by the program PHRED, we compared the actual and predicted error rates for six different cosmid- or BAC-sized projects that were produced by six different large-scale sequencing centers in the United States. All of these six projects used four-color fluorescent sequencing machines; however, the DNA preparation methods, sequencing enzymes, fluorescent dyes and chemistries, and gel lengths varied significantly between the six groups. Table 1 gives an overview of the sequencing projects analyzed. Table 2 lists the different methods used.

RESULTS

Error Rate Prediction Accuracy for Six Projects

A comparison of actual and predicted error rates for the six projects in this study is shown in Table 3.

[1]E-MAIL peter.richterich@genomecorp.com; FAX (781) 893-9535.

8:251 259 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

Table 1. Summary of Data Sets

Project	Reads	Aligned bases	Average aligned read length
A	455	416,214	915
B	1277	871,230	682
C	1065	603,655	567
D	834	414,595	497
E	1638	1,149,209	702
F	1885	907,796	482
Total	7154	4,362,699	610

The results indicate that PHRED is very successful in identifying bases with low error probabilities. For example, the 1.28 million bases with quality scores of 4–12 (corresponding to error probabilities between 39.8% and 6.3%) contain a total of 187,926 errors. In contrast, the 1.44 million bases with quality scores between 33 and 42 (corresponding to error probabilities between 0.05% and 0.006%) contain only 237 errors, which translates into a 790-fold lower error rate. The trend toward lower error rates can also be observed for each individual project. In most cases, the actual number of errors is close to the predicted error rate. It is also apparent that the actual error rate is typically lower than the predicted error rate.

Both the high overall accuracy and the tendency to slightly overpredict errors are confirmed by statistical analysis, as shown in Table 4. The correlation between predicted and actual error frequencies is excellent for all projects (Spearman correlation coefficient >0.89, $P < 0.0001$). Averaged over all projects, the actual error rate is 84.5% of the predicted error rate; the slope of the relation between predicted and actual error rates differs slightly between projects and ranges from 76.6% to 88.4%. To put these differences between projects in relation, it is worthwhile remembering that PHRED quality scores cover a wide dynamic range: The maximum quality score of 51 corresponds to a 50,000-fold lower predicted error rate than the minimum quality score of 4. Even the relative difference between successive quality is larger than the relative difference in the slopes; for example, a quality score of 10 corresponds to an error probability of 10%, whereas a score of 9 corresponds to an error probability of 12.6%.

A different way of looking at the relation between the actual and predicted error rates is shown in Figure 1. Here, the error rates as a function of the position within all reads in each of the projects, averaged over 50-base windows, is depicted. For all six projects, the predicted error rates are very close to the actual error rates over the entire length of the sequences. Each project has a characteristic distribution of error rates, which differs from each of the other projects. The minimum error rate differs dramatically between projects. The best projects achieve raw error rates of 0.23%–0.36% in the best region of the sequence read, typically from base 150 to 200. The worst project in the data set had an ~10-fold higher error rate of 2.58%.

Toward the end of sequence reads, the error rates increase and start to exceed 10% between bases 300 and 700. In projects that used mainly short gels (e.g., projects D and F), this increase begins sooner, whereas projects that use longer gels show a markedly longer stretch of low error rates (e.g., projects A and B).

Table 5 summarizes key results for the six projects. The first four projects have similar minimum and average error rates. However, the length of the region where the error rate is below 5% differs significantly, from 403 to 682 bases. The project with the shorter low error rate regions contained larger portions of reads generated on short gels, whereas projects A and B were run exclusively on long gels (ABI373 stretch or ABI377 sequencers). Other factors contributing to differences between the first four projects were differences in sequencing chemistries, production scale, and electrophoresis conditions and machines.

Project E and, in particular, project F, had significantly higher error rates than the first four projects. In projects E and F, every sequence generated for the project had been included in the data set, whereas the other four projects had eliminated some "bad" sequences through manual or auto-

Table 2. Overview of Sequencing Methods Used in the Different Projects

Template DNA	single-stranded M13, double-stranded plasmids
Sequencing enzymes	Sequenase, *Taq*, KlenTaqTR, AmpliTaq FS
Sequencing chemistries	Dyes primer (two different dyes chemistries), dye terminator
Sequencing machines	ABI 373, ABI 373 stretch, ABI 377
Gel length	Only short gels, only long gels, mixes of short and long gels

Table 3. Comparison of Predicted and Actual Error Rates for Six Different Sequencing Projects

Project	Quality score	4–12	13–22	23–32	33–42	43–51
A	aligned bases	119,246	75,293	70,391	144,876	73,234
	expected errors	20,256	2,064	172	37	1
	actual errors	16,784	1,758	127	17	1
B	aligned bases	182,034	137,940	181,998	399,690	140,176
	expected errors	29,953	3,704	410	102	3
	actual errors	26,038	2,536	287	35	0
C	aligned bases	139,345	131,419	151,197	292,070	68,529
	expected errors	22,277	3,411	357	74	2
	actual errors	16,670	1,513	194	26	3
D	aligned bases	103,898	68,995	68,613	153,730	111,752
	expected errors	16,880	1,919	168	38	3
	actual errors	14,495	1,924	146	59	2
E	aligned bases	378,755	217,438	167,968	392,717	144,313
	expected errors	63,947	6,336	418	95	4
	actual errors	55,968	6,516	355	67	5
F	aligned bases	359,809	136,688	98,840	64,035	5,130
	expected errors	66,938	4,079	256	23	0
	actual errors	57,971	3,856	332	33	1
All	aligned bases	1,283,087	767,773	739,007	1,447,118	543,134
	expected errors	220,252	21,513	1,781	370	13
	actual errors	187,926	18,103	1,441	237	12

matic inspection. After eliminating <10% of the worst sequences in project F, the error rates for the remaining sequences were comparable to those of the first four projects. In contrast, project F showed a much more uniform distribution of sequence quality.

The last column in Table 5 shows the average number of bases with an estimated error probability of at most 0.1%, which is equivalent to a quality score of at least 30. The count of such "very high-quality" bases is a good indicator of sequence quality, both for individual sequences and, when aver-

Table 4. Summary of Statistical Analysis Results

Project	Spearman ρ	$P > \|\rho\|$	Slope	*t* ratio	$P > \|t\|$
A	0.9646	<0.0001	0.818	75.1	<0.0001
B	0.9890	<0.0001	0.874	98.2	<0.0001
C	0.9846	<0.0001	0.766	71.6	<0.0001
D[a]	0.8692	<0.0001	0.855	68.3	<0.0001
E	0.9956	<0.0001	0.884	144.3	<0.0001
F	0.9968	<0.0001	0.865	151.6	<0.0001
All	0.9964	<0.0001	0.845	174.5	<0.0001

[a]In project D, the Spearman correlation coefficient ρ was artificially low as only very few bases (10) bases had a quality score of 5, and none of these bases contained an actual error (expected: 3.16 errors). Exclusion of this quality score gave a Spearman correlation coefficient of 0.9786 ($P < 0.0001$). The frequencies in the slope calculations were weighed by the number of bases at any given quality score and, thus, were not sensitive to such small sample distortions (see Methods).

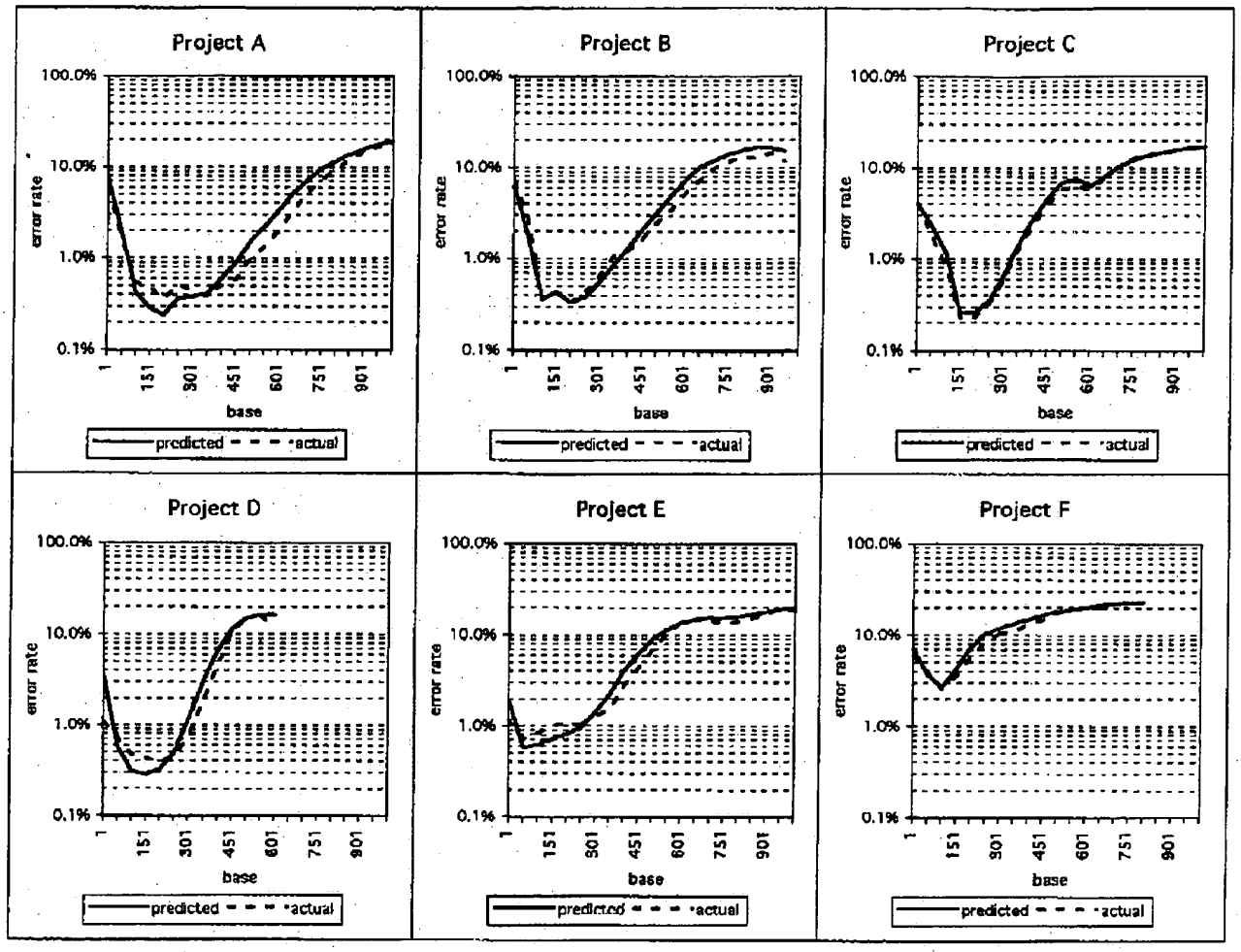


Figure 1 Actual and predicted error rates in six different sequencing projects. Actual error rates and predicted error rates in 50-base windows over the length of the sequence reads, averaged over all reads that could be aligned to the consensus sequence by CROSS_MATCH, are shown. The numbers on the *x*-axis show the first base in a given 50-base window.

aged over all sequences in a project, as an indicator for the entire project. Compared to the estimated error rates, the count of very high-quality bases is less prone to distortions from a small number of low-quality reads, as the data for project E demonstrate.

Prediction Accuracy for Data Subsets of Different Quality

The quality of sequences within any given project can vary substantially, and the use of predicted error rates has the potential to be a powerful tool for qual-

Table 5. Comparison of Key Results for Six Different Sequencing Projects

Project	Actual minimum error rate (%)	Actual average error rate (%)	Length of <1% error region	Length of <5% error region	Average bases with *P*(error) <0.1%
A	0.36	3.6	422	682	468
B	0.34	2.8	274	567	395
C	0.23	2.4	291	479	348
D	0.39	3.1	300	403	294
E	0.71	4.7	129	464	317
F	2.58	9.2	0	162	79

ity analysis and control in large-scale DNA sequencing projects. To analyze how accurate PHRED error estimates are for different quality sequences within the same sequencing project, we subdivided a data set into four quartiles, based on the number of very high-quality bases in each sequence (see Methods). The comparison of actual and predicted error rates is shown in Figure 2.

When measured by the error rate in the best region of a sequence, the data quality in the different quartiles varies >100-fold between the best and the worst 25% of the sequences. The best quartile showed ~0.03% error for >100 bases, whereas the error rate in the worst quartile always exceeded 5%. In quartiles 2 and 3, the predicted error rates match the actual error rates very closely. In the best and worst quartiles, PHRED's accuracy was somewhat lower from base 100 to 500. In the best sequences, PHRED's error estimates were about twofold too high; in the worst sequences, the error estimates were too low, again by a factor of 2. This underprediction of errors can be partially explained by the fact that PHRED gives ambiguous base calls (*N*'s) a quality score of 4, corresponding to an error probability of 39.8%; however, *N*'s will always show up as an actual error. Even in the worst and best quartiles, however, the predicted error rate curves are very similar to the actual error rate curves.

The results shown in Figure 2 also demonstrate that the count of very high-quality bases, or bases with an estimated error probability of at most 0.1%, can be used effectively to characterize the overall quality of a sequence read. Sorting the sequence reads into quartiles based on the number of very high-quality bases worked well, as shown by the >100-fold difference in the minimum error rate between the first and the fourth quartile.

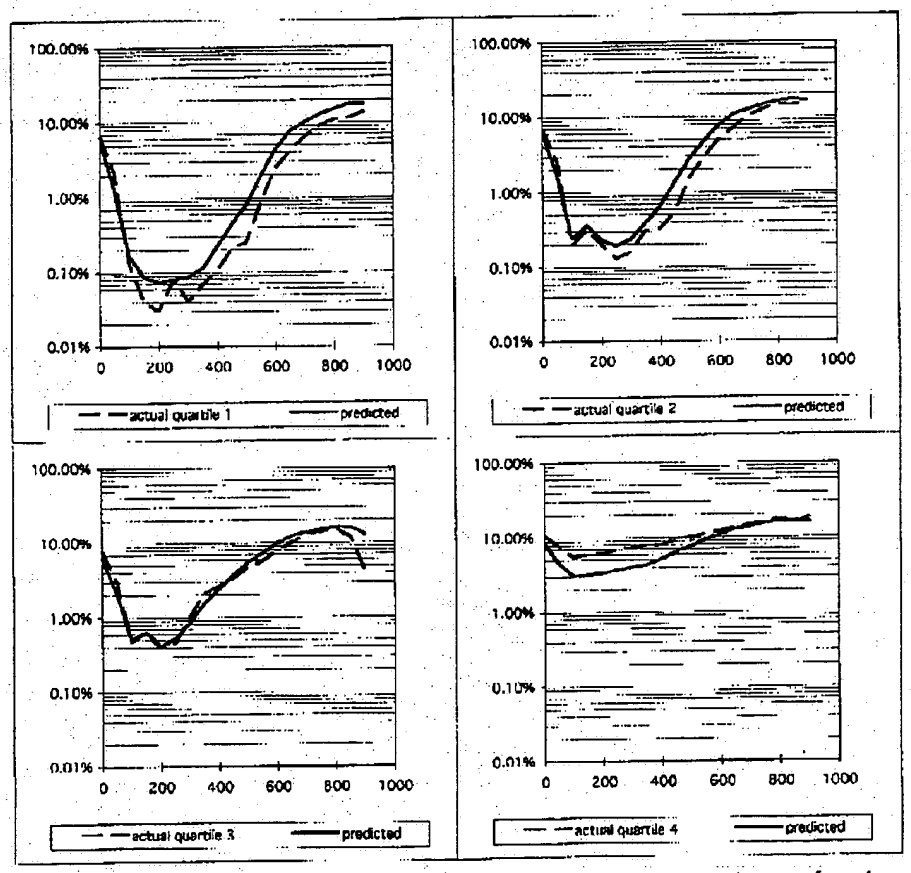


Figure 2 Actual and predicted error rates in different quality subsets of project B. Sequence reads were sorted by the number of bases with a predicted error rate of at most 0.1% (very high-quality bases), and assigned to quartiles, with quartile 1 corresponding to the highest numbers. Actual and predicted error rates for all sequences in each subset were calculated as in Fig. 1. Note that a number of sequence reads that had been rejected because of too low quality were added back to the data set for illustrative purposes, all of which are in quartile 4. These sequences were not included in the data sets used to generate Figs. 1 and 3 and Tables 1 and 3.

Other methods to characterize the overall quality of individual reads based on PHRED quality scores can give similar results. For example, counting bases above a minimum quality threshold anywhere in the range of 20–40 gave similar results for most data sets (not shown), and such counts are used by a number of different laboratories as quality measures. Alternatively, the quality values can be converted to error probabilities and averaged to give the predicted error rate for the trace, or summed to give the total predicted number of errors in a trace. However, such averages and totals can sometimes give a misleading picture, as the following example illustrates. Assume that two sequence reads have very similar quality in the alignable part of the read but that one of the two sequences was run much longer and

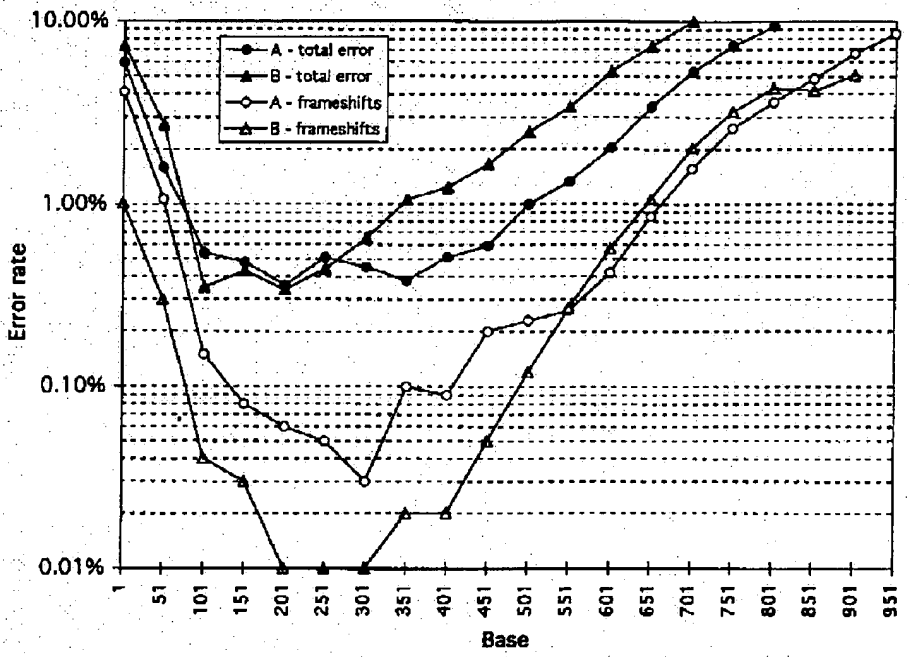


Figure 3 Actual frameshift and total error rates for projects A and B. To calculate frameshift error rates, only insertions and deletions were counted. Mismatch errors, which account for the vast majority of errors after base 150, were included only in the total error count. Note that project B (▲,△) has a slightly similar or slightly higher total error rate compared to project A (●,○) but only about one-third as many insertions and deletions up to base 500. For both projects, the frameshift error rate in the raw data is <1 in 1000 for >300 bases, and ≤1 in 10,000 for >100 bases in project B.

therefore contains a longer unalignable "tail" of very low-quality bases. When calculating the average error rate for these two sequences, the second sequence will have a much higher average error and, therefore, appear to be of lower quality. In contrast, the counts of very high-quality bases for both sequences will be very similar, as the unalignable tails contain few, if any, high-quality bases. Therefore, counts of bases above a high enough quality threshold will give a more robust and clearer picture of trace quality.

Frameshift Error Rates for Different Sequencing Chemistries

Depending on how biologists use DNA sequences, knowledge about total error rates in raw sequences may or may not be sufficient. For example, frameshift errors in coding sequences will generally lead to incorrectly predicted open reading frame, whereas mismatch errors will do so only if the mismatch introduces a stop codon or a new splice site. At the time of this writing, PHRED did not differentiate between mismatch and frameshift errors, but only estimated total error rates. This might occasionally lead to questionable conclusions, as the results shown in Figure 3illustrate.

Figure 3 shows the total actual error rates and the frameshift error rates for two projects, A and B. The total error rates for both projects are similar for up to 350 bases; after 350 bases, project B has a somewhat higher total error rate. However, examining the frameshift error rate gives rise to a different picture: from base 1 to 500, project A has approximately four times as many insertions and deletions as project B. This difference in frameshift error rates can be explained by the sequencing chemistries that were used in the two projects. Project B, with the lower frameshift error rate, used only dye terminator chemistry, which is known to eliminate band spacing artifacts from hairpin structures ("compressions"). Project A, on the other hand, used dye primer chemistry, which is more prone to insertion and deletion errors from mobility artifacts, for most sequencing reactions.

DISCUSSION

As large-scale DNA sequencing has become a more routine and common process, the traditional methods for assessing sequence quality have become unsatisfactory. In projects like single-pass cDNA sequencing, it is not possible to calculate and compare error rates after finishing a sequence, as finishing never takes place. Even when a comparison between raw and finished sequence can be done, the time delay between raw data generation and quality assessment is often large. This delay makes it difficult to improve ongoing projects, and it sometimes makes it impossible to capture problems early on. Some immediate quality feedback can be reached by including known standard sequences for quality control. However, this approach can be costly, and it fails when error profiles differ between standard and unknown sequences.

In contrast to these traditional methods to assess sequence accuracy, direct estimation of error

rates in raw sequence data would enable immediate quality control and feedback. Accurate, base-by-base estimates of error probabilities could also increase the utility of single-pass sequences significantly, allow efficient comparison and optimization of different sequence chemistries, and enable the development of better software tools for sequence assembly and analysis.

The critical question for any error rate prediction tool is how accurate are the error rate estimates, in particular if different sequencing methods and chemistries are used? The results presented herein provide an answer to this question for the program PHRED, as well as clues where further development would be useful. As shown in Tables 3 and 4 and in Figure 1, the agreement between predicted and actual error rates was very good in each of the six different projects analyzed. The observed high level of prediction accuracy in all of these projects is almost astonishing if one takes into account that actual errors are binary (a base is either correct or wrong), whereas predicted error rates are probabilities on a scale from 0.0 to 1.0. The observed tendency to overpredict error rates can be at least partially explained by the "small sample correction" that was used in the derivation of threshold parameters for quality scores (Ewing and Green 1998). For most practical applications, such a somewhat conservative estimation of quality scores is tolerable or even desirable. Overall, the results clearly show that error probabilities given by PHRED accurately describe raw sequence data quality.

In judging the usefulness of predicted error probabilities, it is important to know how differences in sequencing methods will influence the prediction accuracy. For example, the larger variation in peak heights tends to be larger in dye terminator sequencing than in dye primer sequencing, and different sequencing enzymes are known to produce different specific height variation patterns. Any estimation of error probabilities that takes the peculiarities of a specific sequencing chemistry into account would therefore be expected to be less accurate for different chemistries.

The projects included in this study were specifically chosen to provide an initial answer to the question of how generally useful PHRED quality scores are. These projects represent the vast majority of different multicolor fluorescent sequencing methods used in the last 3 years: different template DNAs and DNA preparation methods, different enzymes, gel lengths, run conditions, and different fluorescent dyes. The data also include a considerable spread in data quality, both between projects and within individual projects. None of the projects analyzed here were included in PHRED's training set, and just one of the six laboratories that contributed data to this study also contributed data to the training data sets. One of the projects in this study consisted entirely of dye terminator sequences, which presented only a small fraction of the sequences in the test data set. Another project exclusively used a set of fluorescent dyes different from those used in the training sets. Each project differed from the other projects in this study in at least one, and typically many, experimental aspects like template preparation, sequencing enzymes, gel run conditions, and so forth. Despite these differences, the accuracy of error rate predictions was very similar for all projects.

Our results justify some optimism about the accuracy of PHRED quality scores for minor changes in sequencing technology, for example, sequences generated by new enzymes and fluorescent dyes. Initial studies showed that PHRED quality scores were also accurate for sequences produced by multiplex sequencing with radioactive detection (P. Richterich, unpubl.). However, we also observed two effects that can invalidate PHRED quality scores during these studies. First, sequences generated by chemical sequencing gave too low quality scores at mixed (A + G) reactions. Because secondary peak height is one of the parameters used in the error rate predictions, this is not surprising. Another potential source of error is high-frequency noise in the trace data. With such data, PHRED occasionally underestimated the band spacing by a factor of 2 or more, which resulted in incorrect base calls and quality scores. By applying simple smoothing algorithms to data with high-frequency noise, these problems could typically be resolved. Similar steps may be necessary to obtain accurate PHRED quality scores on data that have been generated by different sequencing instruments or preprocessed by different software.

Accurate quality scores can have a major impact on how sequences are used downstream from the sequence production process. In traditional sequencing projects where the goal is complete coverage at a final error rate below (e.g.) 1 in 10,000, the accuracy goals can be reached with single sequence reads as long as the quality scores are at least 40 (however, other potential problems like clone instability may make higher coverage advisable). Interesting questions arise as to how individual read quality contributes to project quality, or the error rate of the "final" sequence. Under the assumption that errors between different sequence reads are

completely independent, one could argue that two reads with a quality score of 20 (error probability of 1 in 100) are just as valuable as one sequence with a quality score of 40 (error probability of 1 in 10,000). However, although a single sequence stretch with quality levels above 40 would give a final sequence with an error rate of <1 in 10,000, assembling a consensus from two sequences with quality scores of 20 (1% error rate) could lead to one of two results: If the errors were completely random, the consensus sequence would be ambiguous at 2% of all locations; if the errors were completely localized, for example, because of reproducible compressions, the consensus sequence would have one "hidden" error every 100 bases. Typically, consensus sequences derived from low-quality sequences will have both kinds of problematic regions. Increased coverage can rapidly eliminate the random errors; however, increased coverage does not resolve errors from systematic sources. Manual examination of such problem areas is generally required; such "contig editing," however, tends to be time consuming, requires highly trained personnel, is an obstacle toward complete automation of DNA sequencing, and sometimes fails to eliminate all errors. This leads to the somewhat counterintuitive conclusion that the practical value of increasing sequence quality can be even higher than indicated by the quality scores: One sequence of average quality above 40 can be "worth" more than two sequences of average quality 20.

Another application of DNA sequencing where high quality can be of disproportionately high value is the search for mutations in genomic DNA. In low quality sequences, secondary peaks and low resolution often complicate the identification of heterozygous mutations. In regions of higher sequence quality, such secondary peaks are smaller or absent and peaks are better resolved. Therefore, both false-positive and false-negative errors can be significantly reduced in high-quality regions. Tools like PHRED, which can accurately measure sequence quality from trace data, can be of twofold value for mutation detection. First, base-specific quality scores can allow optimization of sequencing methods and strategies for mutation detection. Second, the quality scores can be used to evaluate the usefulness of individual sequence reads for mutation detection (e.g., by discarding reads below minimum thresholds), and they can guide software that automatically detects mutations.

The ability to predict error rates in a highly accurate fashion is likely to have a major impact in applications like those described above. PHRED is the first widely used program that accurately predicts base-specific error probabilities. However, the algorithm for determining quality values has been described (Ewing and Green 1998), and it should be straightforward to implement similar quality values in other base-calling programs. Furthermore, an extension of the approach developed by Ewing and Green should be possible. For example, differentiation between mismatch and frameshift errors would enable better comparisons of sequencing methods with similar total error rates but different frameshift error rates. Several groups have described efforts to calculate separate probabilities (or "confidence assessments") for mismatch errors and frameshift errors (Lawrence and Solovyev 1994; Berno 1996). Their results demonstrated that different approaches to error type characterization are feasible and promising. Implementation of such error type predictions in other programs similar to the way PHRED uses quality scores would enable better method assessments, benchmarking, and production quality control, and could have a significant impact on downstream uses of DNA sequence information.

METHODS

Data Sets

For one project, sequence raw data in the form of ABI trace files were downloaded from a public FTP site. Sequence data for the five other projects were kindly provided by five different large-scale sequencing groups. Table 1 gives a summary of the six projects, and Table 2 gives an overview of the different sequencing methods used in the projects. The projects differed in the amount of prescreening of data that had been done, reflecting different approaches to quality control in different laboratories. In two projects (B and C), different software programs had been used to identify and eliminate low-quality sequences. One project (F) included all data files generated, whereas the other three projects had excluded "failed lanes."

Comparison of Actual and Predicted Error Rates

The sequences for all traces in each project were recalled using the program PHRED (v. 961028). Next, sequences in each project were assembled with PHRAP (P. Green, unpubl.). Slightly different methods were chosen for the statistical and graphical evaluation of the error rate prediction accuracy. In the statistical evaluation, only the longest contig produced by PHRAP was considered. The tables of aligned bases and observed discrepancy counts for

each quality score were taken from the PHRAP output and analyzed as follows. The expected number of discrepancies (E) at each quality score (q) was calculated by multiplying the number of aligned bases (N) with the error probability corresponding to the quality score: $E = N\,10^{-0.1q}$. The Spearman ranking coefficients were calculated by comparing the expected and observed error frequencies. To obtain the quantitative relation between the expected and observed error rates over the entire range, a least-squares fit between the observed and expected rates was performed, with the intercept set to zero and the number of aligned bases at each quality score used as weights.

For a graphical comparison of estimated and actual error rates in 50-bp windows, the following steps were taken. For two of the projects, the consensus sequence was retrieved from public databases. For the four other projects, the DNA sequence and quality information were used by the program PHRAP to assemble consensus sequences for each of the projects. The individual reads were aligned to the consensus sequences of the longest contig, using the program CROSS_MATCH (P. Green, unpubl.), after removing single-coverage regions from the ends of the consensus sequence. CROSS_MATCH uses an implementation of the Smith Waterman algorithm to generate alignments that typically do not include the ends of sequences, where disagreements are commonly due to vector sequence or low quality sequence.

The quality files generated by PHRED and the alignment summaries generated by CROSS_MATCH were then analyzed as follows. First, the region of each query sequence that had been aligned by CROSS_MATCH was determined. Next, the actual and predicted error rates for the entire aligned part of each individual sequence was calculated. In addition, the average actual and predicted error rates for all alignable sequences together were calculated for windows of 50 bases in length. To calculate the predicted error rate, the quality scores q determined by PHRED at each base were converted to error probabilities as described above (Ewing and Green 1998).

Subdividing Data Into Subsets Based on Data Quality

To examine the accuracy of PHRED quality scores for data subsets of different quality within a project, the following approach was taken. For all sequence reads in project B, the number of bases with a quality score of at least 30 in each sequence was determined (bases with quality scores of at least 30 were called very high-quality bases, or VHQ bases). Sequences were sorted in descending order based on the number of very high-quality bases, and divided into four quartiles. Accordingly, quartile 1 contained 25% of sequences with the highest number of very high-quality bases, and quartile 4 contained the "worst" sequences. To illustrate the prediction accuracy in data with relatively high error rates, sequences from project B that had been "discarded" because they had not met the minimum quality criteria were added back to the data set. The sequences in each quartile were compared to the consensus sequences that had been generated using the entire data set, as described above for the graphical comparison.

Determining Actual Frameshift Error Rates

The calculation of actual frameshift error rates in the raw sequence data was performed using CROSS_MATCH, similar to the procedure described above for total error rates, except that only insertion and deletion errors were counted. Because PHRED does not give separate frameshift error estimates, a comparison of predicted and actual frameshift errors is not possible.

ACKNOWLEDGMENTS

I thank the participating laboratories for contributing their data, Dr. Josée Dupuis for help with the statistical analysis, and Dr. Phil Green for helpful discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bernd, A.J. 1996. A graph theoretic approach to the analysis of DNA sequencing data. *Genome Res.* **6:** 80–91.

Bonfield, J.K. and R. Staden. 1995. The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Res.* **23:** 1406–1410.

Churchill, G. and M.S. Waterman. 1992. The accuracy of DNA sequences: estimating sequence quality. *Genomics* **14:** 89–98.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using *phred.* II. Error probabilities. *Genome Res.* (this issue).

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using *phred.* I. Accuracy assessment. *Genome Res.* (this issue).

Lawrence, C.B. and V.V. Solovyev. 1994. Assignment of position-specific error probability to primary sequence data. *Nucleic Acids Res.* **22:** 1272 1280.

Received October 27, 1997; accepted in revised form February 3, 1998.

GENOME RESEARCH

Volume 8 Number 3
March 1998

Editorial Office

Cold Spring Harbor Laboratory Press
1 Bungtown Road
Cold Spring Harbor, New York 11724
Phone (516) 367-6834
Fax (516) 367-8334
http://www.cshl.org

Editorial/Production

Nadine Dumser, Technical Editor
Kristin Kraus, Production Editor
Cynthia Grimm, Production Editor
Peggy Calicchia, Editorial Secretary